REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY) 28-03-2015	2. REPORT TYPE Final Report	3. DATES COVERED (From - To) 1-May-2014 - 31-Jan-2015

4. TITLE AND SUBTITLE
Final Report: Rapid Characterization of Spider Silk Genes via Exon Capture

5a. CONTRACT NUMBER
W911NF-14-1-0145

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS
Cheryl Y. Hayashi, Matthew A. Collin

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Riverside
200 University Office Building

Riverside, CA 92521 -0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
64829-LS-II.1

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

Spider silks are high-performance materials with an array of potential military and civilian applications. As such, there is persistent demand for the mass production of silks, which requires knowledge of the underlying silk gene sequences. Spidroins (spider fibroins), the most abundant proteins in silks, have repetitive internal regions flanked by non-repetitive amino- and carboxyl-terminal regions. The terminal regions are integral for silk processing and fiber formation, while the repetitive regions are integral for fiber self-assembly and mechanical properties. Full characterization of spidroin genes has been challenging due to their great lengths and inherent 3' bias of cDNA

15. SUBJECT TERMS
aggregate silk glue, de novo assembly, next generation sequencing, silk, spider, spidroin, target capture

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Cheryl Hayashi
UU	UU	UU	UU		19b. TELEPHONE NUMBER 951-827-4322

Report Title

Final Report: Rapid Characterization of Spider Silk Genes via Exon Capture

ABSTRACT

Spider silks are high-performance materials with an array of potential military and civilian applications. As such, there is persistent demand for the mass production of silks, which requires knowledge of the underlying silk gene sequences. Spidroins (spider fibroins), the most abundant proteins in silks, have repetitive internal regions flanked by non-repetitive amino- and carboxyl-terminal regions. The terminal regions are integral for silk processing and fiber formation, while the repetitive regions are integral for fiber self-assembly and mechanical properties. Full characterization of spidroin genes has been challenging due to their great lengths and inherent 3' bias of cDNA-based methods. In this STIR project, new target genome capture and bioinformatic approaches were utilized to rapidly and economically determine genetic blueprints for spider silks. A silk gene bait set was developed and sixteen genomic capture libraries, representing six species, were constructed and sequenced. The results dramatically expand knowledge of spidroin genes, especially for the less well-documented amino-terminal coding regions, and also non-spidroin silk genes, particularly the aggregate silk glues. Many novel silk sequences were obtained, showing the feasibility of target capture for gene discovery. The new silk gene sequences are highly relevant for understanding structure/function of spider silk proteins and recombinant silk production.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received	Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
Matthew A. Collin	0.47
R. Crystal Chaw	0.22
FTE Equivalent:	**0.69**
Total Number:	**2**

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Cheryl Y. Hayashi	0.00	
FTE Equivalent:	**0.00**	
Total Number:	**1**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

Names of Personnel receiving masters degrees

NAME
Total Number:

Names of personnel receiving PHDs

NAME
Total Number:

Names of other research staff

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

See Attachment for Report with Table and Figures.

List of Illustrations

List of Tables

Technology Transfer

Statement of the problem studied

In this project, we pioneered new methods to rapidly and economically characterize genetic blueprints for spider silks. Spider silks are renowned for their impressive physical properties, including high tensile strength, extraordinary toughness and extreme extensibility. Spider silks are also remarkably lightweight and synthesized at ambient temperature. Because silks are primarily composed of protein, they are biodegradable and more eco-friendly than manmade fibers such as nylon and Kevlar, which are industrially manufactured using harsh chemicals and solvents. These advantageous traits are a few of the reasons underlying the extensive research on synthesizing materials that replicate the properties of spider silks.

Spiders produce multiple types of silk, with each silk type having unique properties tailored for specific tasks. For example, the capture spiral of an orb-web is composed of sticky glue (aggregate silk) and a highly extensible filament (flagelliform silk) that work together to arrest flying prey. Specific amino acid sequence motifs contribute to these mechanical properties via the folding and interaction of higher-level structures. Most of the bulk proteins in spider silks are spidroins (contraction of "spider fibroins"). Spidroins have exceptionally high molecular weights (often >300 kDa) and are characterized by short, non-repetitive amino- and carboxyl-terminal regions that flank very long strings of tandem-arrayed repeats. These regions of repeated amino acid sequence motifs are encoded by vast stretches of repetitive DNA, which pose technical challenges for gene cloning and sequencing.

Spidroin and other spider silk genes have largely been characterized through Sanger (traditional) sequencing of cDNAs, and more recently via high-throughput (next generation) RNA-seq methods. Both methods are biased towards reporting the 3' portion of transcripts. Thus, the vast majority of spidroin gene sequences are incomplete, typically encompassing only the portion of 3' repetitive region that immediately precedes the non-repetitive carboxyl-terminal region. This means that little is known about spidroin amino-terminal regions, which are thought to be important for silk synthesis. Furthermore, it has not been feasible to quickly scan an individual spider's genome for silk genes.

In this ARO STIR funded project, we successfully showed that targeted genome sequencing is an effective approach to dramatically expand knowledge about all regions of silk genes (amino-terminal, repetitive, and carboxyl-terminal encoding), and survey the suite of silk genes within a spider's genome, even from species with limited prior genetic information.

Experimental Design

Target Genome Capture Overview

In brief, target genome capture is an approach to subsample whole genomes for regions of interest. For example, target genome capture is used in human genomic studies for resequencing of exons or particular chromosomes. By winnowing the genome for sequences of interest, more efficient use is made of high-throughput sequencing and data analysis. Target genome capture relies on the synthesis of a library of biotinylated "baits" (oligomers) derived from the "targets" (sequences of interest). Genomic DNA libraries are constructed from DNA randomly sheared into appropriately sized fragments. The genomic libraries are hybridized to the bait library, then the "captured" genomic fragments are bound onto streptavidin-coated magnetic beads, followed by a series of washes to remove unbound, non-target genomic DNA fragments. The "captured" genomic DNA fragments are then sequenced on a high-throughput platform.

Study Species and Bait Design

There were two major considerations for the selection of focal species, (1) relevance to the silk research community and (2) number of annotated silk sequences already deposited in the NCBI GenBank. Based on these factors, four species were chosen: *Araneus diadematus* (cross orb-weaver), *Argiope argentata* (silver garden spider), *Latrodectus hesperus* (Western black widow), and *Nephila clavipes* (golden silk orb-weaver). *A. diadematus* and *N. clavipes* silk gene sequences are extensively utilized for recombinant silk production in microbial expression systems. Additionally, *A. diadematus* sequences have been inserted into goats and *N. clavipes* sequences have been inserted into the domesticated silkworm for production in eukaryotic hosts. Finally, there are several complete spidroin gene sequences on GenBank for *A. argentata* and *L. hesperus*.

To assess the robustness of target capture in light of sequence divergence, we also studied *Parasteatoda tepidariorum* (common house spider) and *Steatoda grossa* (false black widow). *P. tepidariorum* and *S. grossa* are members of the Theridiidae (cob-web weaver) family, as is *L. hesperus*, which has the largest diversity of known spidroin and silk associated genes within our dataset. No DNA sequences from *P. tepidariorum* or *S. grossa* were included in our baits. Thus, any silk gene sequences we obtain from these species demonstrate the effectiveness of heterologous target capture.

To develop the bait library, we began by downloading GenBank entries for spidroin and other silk-associated genes from the focal species (*A. diadematus*, *A. argentata*, *L. hesperus*, *N. clavipes*). Because comparatively few genes were known from *A. diadematus*, we also downloaded genes from other *Araneus* species. For each focal species, sequences with at least 98% identity over at least 100 bases were combined into contigs to reduce redundancy. The curated sets of target sequences for the focal species were appended to each other and sent to Agilent Technologies. There, a large pool of 120 bp oligomers were designed and synthesized from the sequences. This was our custom-made bait library.

Library Preparation and Sequencing

We prepared capture libraries for sequencing with two different methodologies, (1) Illumina MiSeq and (2) Roche 454. Both methods perform parallel sequencing of numerous templates, but MiSeq produces massive numbers of shorter reads (<300 bp) via reversible chain termination while 454 yields fewer numbers of longer reads (>600 bp) via pyrosequencing. These contrasting sequencing methods were pursued to empirically determine the effectiveness of long reads to improve silk gene contigs assembled from short reads.

Twelve Illumina libraries were constructed for the four focal species (*A. diadematus*, *A. argentata*, *L. hesperus*, *N. clavipes*) and two divergent species (*P. tepidariorum*, *S. grossa*) using the SureSelect^{XT} Target Enrichment System for Illumina Paired-End Sequencing Library kit (Agilent). Each library was constructed from DNA extracted from a single spider. For each species, two libraries were constructed as biological replicates.

Libraries were uniquely indexed and multiplexed sequenced four at a time at the UC Riverside Genomics Core facility (MiSeq, 2 x 300 bp).

Four Roche 454 libraries were constructed for the focal species. Each Roche library was constructed from a genomic DNA extraction also used to construct a MiSeq library. Thus direct comparison of sequences yielded from each sequencing method could be made without confounding factors such as allelic variation between individuals. Genomic DNA library construction and hybridization to capture baits were done with the SureSelect Target Enrichment System for Roche 454 GS FLX and GS Junior Sequencing Platforms kit (Agilent). Because unique indexing of libraries was not supported by the Agilent kit, rather than make a multiplex cocktail, each library was confined to its own quadrant of a 454 plate. Roche 454 libraries were sequenced at the University of Arizona Genetics Core facility with GS FLX+ chemistry.

Read Processing and Assembly

Pre-assembly read processing for all Illumina datasets followed a protocol designed to minimize incorporation of low quality bases into our results. First, adapters and barcodes were removed, and reads containing any ambiguous bases (basecalls other than A, C, G, T) were removed. Next, reads were scanned for quality and average base composition at each position. Based on these scans, the first 6-10 bases were removed from the 5' end of each read due to lower quality and skewed base composition, attributed to artifacts from the library construction steps (e.g., A-tailing). Also based on the quality statistics, the 3' ends of the reads were trimmed from the first position (reading from 5' to 3') with a first-quartile quality score that dropped below Q25. Reads were then selectively deleted based on quality score over the entire length of the read. Reads with < 85% of the bases Q25 or better were removed.

The Trinity assembler was used to construct contigs from the Illumina datasets. Contigs relevant to particular analyses were identified by matching to the reference sequences from GenBank using Geneious v. 6.16 (Biomatters). Several preset stringencies were used. Relevant contigs were also identified with blastx searches to a database containing either spidroin termini or non-spidroin silk protein sequences (i.e., avoid cluttering analyses with spidroin repetitive regions). The contigs with hits were then confirmed through blastx searches against the NCBI nr database. These multiple approaches were pursued to maximize the number or silk contigs discovered.

Read filtering and contig assembly of Roche 454 datasets were performed using Newbler (Roche). Newbler was developed to account for differences in read quality and length due to pyrosequencing. Newbler assembles contigs by overlapping alignment of reads, taking into account base quality scores. This method is computationally unfeasible for Illumina datasets, which have two to three orders of magnitude more reads. Relevant 454 contigs were identified with the same approach described above for MiSeq contigs.

De novo assemblers have difficulty with repetitive DNA sequences. Thus, we developed an algorithm to determine the most prevalent repeats for a given spidroin rather than attempt to accurately assemble tandem repeats. The multistep process begins with trimmed and filtered reads. These reads are mapped to known spidroin repetitive regions using Bowtie. The matching reads are iteratively assembled using a position-weight matrix to produce a model repeat for a specific spidroin.

Results

The twelve Illumina MiSeq libraries resulted in a total of 40Gb of raw sequence data, from 136 million reads that were 250-300 bp in length. After trimming and filtering, 12 Gb in 86 million reads were used for data analyses (Table 1). Separate Trinity assemblies were done for each MiSeq library. The number of contigs assembled from each library ranged from 29,000 to >200,000, with mean contig lengths of 297-422 bp. Many of the contigs had significant blastx scores (<1e^{-5}) to our target sequence database. *L. hesperus* assemblies had the largest proportion of "on-target" longer contigs while *P. tepidariorum* had the lowest (35% and 5%, respectively). Because *de novo* assembly programs have difficulty with building contigs from repetitive DNA, contigs derived from spidroin genes tended to be short (truncated). Despite this technical challenge, numerous contigs >1000 bp were assembled.

Roche 454 sequencing generated reads that were 600-1000 bp, about two to three times the length of MiSeq reads. The four 454 capture libraries produced a total of 202 Mb of sequence, from >347,000 reads (Table 1). Libraries were individually assembled with Newbler and the assemblies ranged in size from ~1,100 to 5,300 contigs, with mean contig lengths of 634 to 933 bp. Additionally, a substantial percentage of contigs (up to 50% for *L. hesperus*) had significant blastx scores to the target sequences.

Table 1. Capture Library Sequencing and Assembly Statistics. N50 is the weighted median length of contigs in the assembly. L50 is the number of contigs that account for more than 50% of the assembly.

Library	no. reads	no. bases	no. contigs	N50	L50
MiSeq Target Species					
Argiope argentata 1	10,027,726	2,968,130,541	77,143	382	296
Argiope argentata 2	8,802,962	2,625,463,693	45,934	293	228
Araneus diadematus 1	12,467,102	3,705,738,102	58,674	405	287
Araneus diadematus 2	13,042,852	3,880,912,146	48,947	408	293
Latrodectus hesperus 1	10,378,966	3,084,143,178	29,062	386	291
Latrodectus hesperus 2	12,450,472	3,698,363,421	201,651	275	232
Nephila clavipes 1	11,352,644	3,397,699,458	36,712	434	297
Nephila clavipes 2	10,522,816	3,127,276,348	118,435	295	227
MiSeq Non-Target Species					
Parasteatoda tepidariorum 1	13,103,714	3,930,103,629	35,730	402	267
Parasteatoda tepidariorum 2	10,651,948	3,183,572,017	214,274	323	272
Steatoda grossa 1	11,954,402	3,572,321,038	60,422	295	259
Steatoda grossa 2	11,922,590	3,558,637,440	186,711	304	265
454 Target Species					
Argiope argentata 2	89,478	49,677,649	1,093	1,032	269
Araneus diadematus 1	59,593	33,713,576	1,883	840	494
Latrodectus hesperus 2	63,882	37,315,641	1,347	826	343
Nephila clavipes 1	134,725	81,478,923	5,302	1,245	1,477

Accomplishments

Through the use of capture libraries, we recovered all of the previously known target sequences from each of the four focal species and also gained novel sequences for each species (Figure 1). The newly characterized sequences represent an increase in knowledge of between 4% (*L. hesperus*) to 775% (*A. diadematus*). The 4% *L. hesperus* gain was from two sequences that are putatively flagelliform spidroin amino-terminal regions. The enormous 775% *A. diadematus* gain is due to this species having only four target sequences on GenBank, but our capturing of 28 new sequences with the baits designed from the other species.

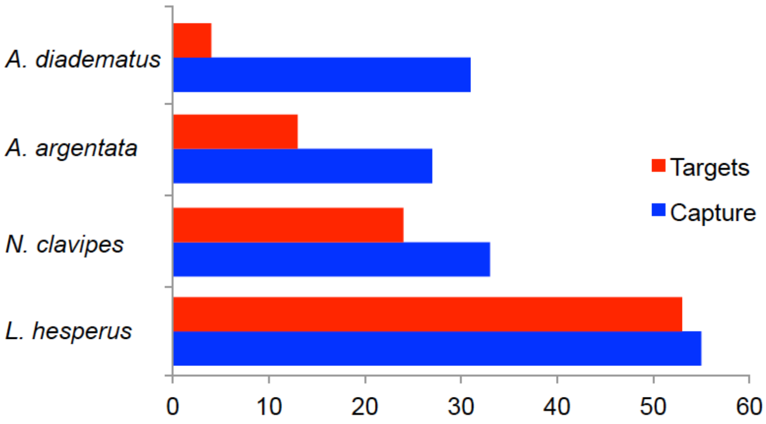
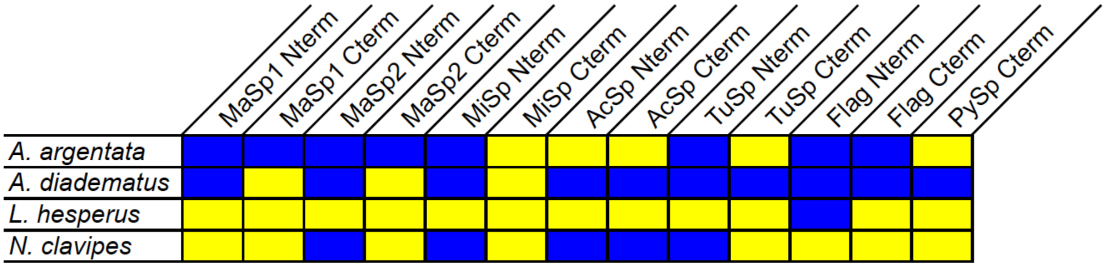


Figure 1. Comparison of unique targets per species to unique captured sequences. To be considered "unique," a sequence must vary sufficiently from other sequences. Thus, sequences with ≥ 98% identity over at least 100 bp were combined into a consensus sequence and counted once.

For each of the four focal species, we obtained between 27 to 31 terminal sequences across seven spidroin types (Figure 2). In total, more than 120 amino- and carboxyl-terminal regions for the seven silk types, were classified across the six species. For well-characterized species, such as *L. hesperus*, terminal region contigs closely matched the GenBank sequences and we could determine the minimum number of spidroin gene copies in a black widow genome. In less well-characterized species, such as *A. diadematus*, our data expand the number of described spidroins greater than 7 fold, with numerous allelic variants. Moreover many of the terminal sequences are amino-terminal regions, which are the most poorly known parts of spidroins. Amino- and carboxyl-terminal coding regions are both important because they can be used to amplify complete silk genes by providing priming sites that span start to stop codons. The newly discovered amino-terminal regions could also be used to study *in vivo* silk production through labeling the spidroins or silencing specific gene to determine the effect on silk production and fiber mechanical properties.

Target capture enabled the rapid characterization of spidroins from species with limited prior knowledge of silk genes. For example, only four spidroin cDNA sequences have been published from *A. diadematus*. Despite this small number, *A. diadematus* spidroin sequences are used for recombinant silk production in microbial and eukaryotic hosts. With target capture, we increased the number of spidroin terminal regions for *A. diadematus* to ten, including amino-terminal regions for five spidroins (Figure 2). We obtained similar numbers of terminal regions from each capture library. For the better-characterized species, the same set of spidroins was assembled, demonstrating that target capture can be effective for identifying the genetic basis for silks with exceptional properties within a species. Additionally, for the non-target species, *S. grossa* and *P. tepidariorum*, we recovered 27 spidroin sequences despite having no baits specific to these species. Thus, target capture can also be used to discover spidroins from previously unstudied species.



	MaSp1 Nterm	MaSp1 Cterm	MaSp2 Nterm	MaSp2 Cterm	MiSp Nterm	MiSp Cterm	AcSp Nterm	AcSp Cterm	TuSp Nterm	TuSp Cterm	Flag Nterm	Flag Cterm	PySp Cterm
A. argentata	🟦	🟦	🟦	🟦	🟦	🟨	🟨	🟨	🟦	🟨	🟦	🟦	🟨
A. diadematus	🟦	🟨	🟦	🟨	🟦	🟨	🟦	🟦	🟦	🟦	🟦	🟦	🟦
L. hesperus	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟦	🟨	🟨
N. clavipes	🟨	🟨	🟦	🟨	🟦	🟨	🟦	🟦	🟦	🟨	🟨	🟨	🟨

Figure 2. Spidroin terminal region contigs assembled from target capture reads for the focal species. “Nterm” = amino-terminal region, “Cterm” = carboxyl-terminal region, yellow boxes = contigs for which there was previously known sequence, blue boxes = novel termini discovered from this study. No PySp Nterm contigs were assembled (there were no baits for PySp Nterm sequence).

We found that assembly with the longer Roche 454 reads resulted in increased length of spidroin contigs. When comparing the 454 contigs to MiSeq contigs, an equal diversity of spidroin termini was represented despite the vastly fewer 454 contigs (Table 1). However there were fewer allelic/loci variants present among the 454 contigs, which can be attributed to the fewer number of reads from the 454 sequencing platform compared to the MiSeq. Comparing N50 values, the average 454 assembled contig is longer than the average MiSeq assembled contig. Contigs covering the same genomic regions from 454 and MiSeq assemblies were aligned with the corresponding target sequence (e.g., *L. hesperus MaSp2*; Figure 3). The 454 assembled contig for *MaSp2* from *L. hesperus* is longer than the best matching target *MaSp2* sequence fragment and is identical in sequence to two non-overlapping MiSeq contigs. The same pattern is observed with the *N. clavipes* ASG2 contigs (see below). The longer 454 contigs expanded knowledge of amino-terminal coding regions and showed putative inaccuracies in the ASG2 sequence on GenBank.

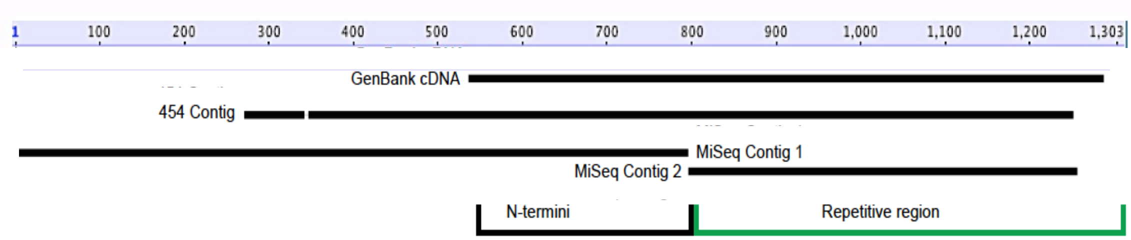


Figure 3. Comparison of *L. hesperus* MaSp2 contigs assembled from Roche 454 and Illumina MiSeq capture library reads to a published MaSp2 cDNA with amino-terminal region coding sequence (GenBank EF595248). The 454 Contig encompasses the amino-terminal region and the immediately adjacent repetitive region, whereas the MiSeq assembly resulted in two separate contigs for the same region. Scale bar in bases.

We have begun developing a bioinformatic method to utilize the millions of spidroin repetitive region reads to allow comparison of the most prevalent repeats across species. As a start, we are analyzing aciniform spidroin (AcSp1) repeats. Thus far, our algorithm successfully determined the AcSp1 repeat for three focal species and *S. grossa* of the non-target species. The algorithmically derived repeat for *L. hesperus* is 1128 bp and *A. argentata* 612 bp, both matching the published data in length and nucleotide sequence. We also determined the AcSp1 repeat for *S. grossa* (1104 bp) and *A. diadematus* (1247 bp), which have not been published. We are currently refining the method and applying it to other spidroin repeat types.

In addition to spidroins, we also targeted non-spidroin silk genes with our capture baits. For example, genes for aggregate silk glues 1 and 2 (ASG1 and ASG2) were assembled for all six species that we studied. The published cDNA sequences for ASG1 and ASG2 glues from *N. clavipes* have the unusual feature of a 351 bp repetitive region shared by both sequences that is in one orientation in ASG1 and in the reverse-complemented orientation in ASG2. For all six species in our study, our contigs do not support the existence of a repetitive region shared by ASG1 and ASG2. Instead, we found no evidence for the 351 bp Ser-Gly-Ser repetitive region reported for *N. clavipes* ASG2. We also mapped our paired-end reads from our *N. clavipes* MiSeq libraries to the GenBank ASG2 sequence (EU780015) to determine if any of the 14+ million read pairs mapped to the missing repetitive region. No reads mapped to the Ser-Gly-Ser repetitive region as shown by the read depth graph (Figure 4). Additionally, our results show that there is ~1000 bp of coding sequence preceding the start codon in the GenBank entry for N. clavipes *ASG2*. These findings may impact the development of recombinant silk glues by improving the accuracy and completeness of the *N. clavipes* ASG2 gene sequence and providing multi-species comparative data for ASG1 and ASG2.

Future recommendations

Targeted sequencing through bait capture is a fast and efficient method for the characterization of spidroins and associated silk proteins. Several spidroin termini fragments from *Parasteatoda* (13 contigs) and *Steatoda* (14) were assembled showing the potential of our bait set to recover spidroins from non-target species despite ~60% nucleotide sequence divergence. MaSp1 was the most abundant spidroin captured from *Parasteatoda* (4) and *Steatoda* (5). With further bait optimization, targeted sequencing should accelerate the discovery of spidroins and silk associated sequences.

We recommend that it would be worthwhile to construct a comprehensive bait set from all known spidroin termini to effectively target only terminal regions from a broad range of spider species. Not including repetitive regions into the bait set will reduce the size of the expected genome capture. Yet, enough flanking repetitive sequence adjacent to the terminal regions will be captured to facilitate matching of termini with spidroin type. Smaller genome captures will allow more libraries to be sequenced in a single lane, thereby reducing research costs. Importantly, spidroin terminal regions will be more deeply sequenced, increasing confidence in the completeness of the genome sampling.

Bait capture should improve the utility of new technologies, such as Pac Bio SMRTBell sequencing, Illumina Synthetic Long Reads and Oxford Nanopore MinION, to spidroin discovery. These technologies aim to increase sequence length by using longer (> 8 kb) molecules as starting material, which would be well suited to spidroin genes. The vast amount of repetitive sequence within spidroin genes improves the chances for target capture because numerous baits can bind to a genomic DNA fragment and then sequenced over their entire (or nearly entire) length.

Summary

We showed that our target capture and bioinformatic methods worked well to isolate spidroins and associated silk proteins from spider genomes. Through the use of this methodology we captured previously uncharacterized spidroin terminal regions, repetitive regions, and revised the gene model for ASG2 aggregate glue. Comparative analyses of these sequences can be used to improve production of recombinant spider silk. Furthermore, the newly identified 5' gene sequences can be used to improve silk production by serving of priming sites to clone full-length spidroin genes, baits for future target capture (especially with very long genomic DNA fragments), and for studies of gene function via RNA silencing.

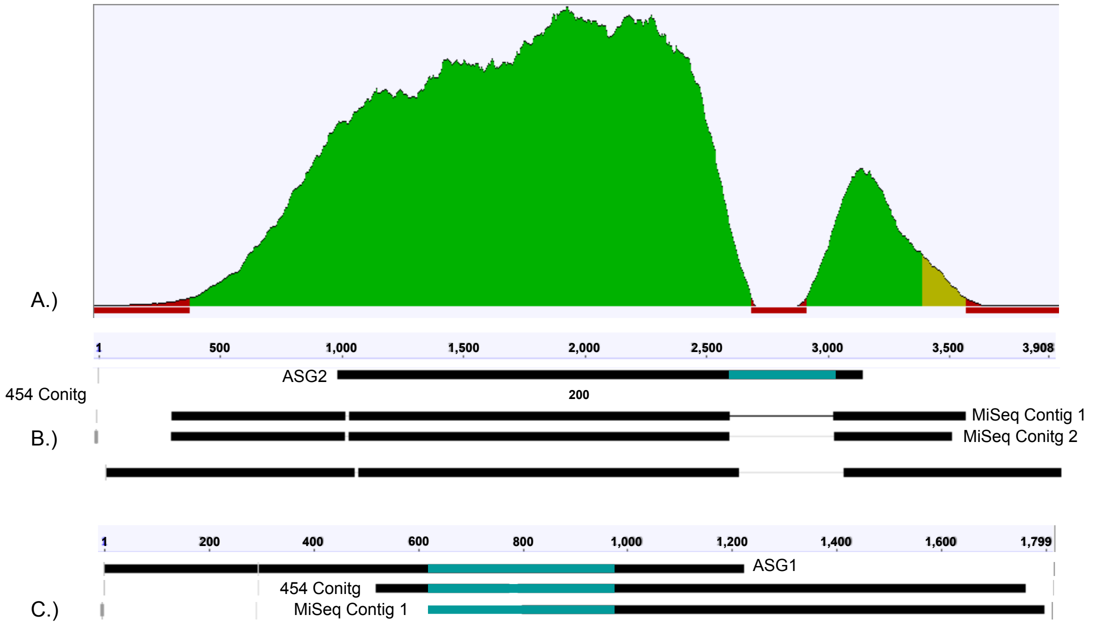


Figure 4. Comparison of target capture contigs and GenBank sequences for *N. clavipes* aggregate silk glue 1 (ASG1) and 2 (ASG2). A.) Depth of coverage graph for paired-end reads mapped to GenBank ASG2 (EU780015), green indicates >50 mapped reads, yellow < 50, and red < 10. No reads mapped to the 351 base repeat region reported in GenBank ASG2. B.) Alignment of GenBank ASG2 with ASG2 contigs from MiSeq and 454 capture libraries. 351 bp repeat region (teal) of the GenBank entry is absent from the contigs. C.) Alignment of GenBank ASG1 (EU780014) with ASG1 contigs from the capture libraries. 351 base repeat region (teal) is present in all sequences, reverse complemented compared to orientation in ASG1. Scale bars in bases